J. SIAM NUMER. ANAL.
Ser. B, Vol. 2, No. 1
Printed in U.S.A., 1964

THE SOLUTION OF ELLIPTIC DIFFERENCE EQUATIONS BY SEMI-EXPLICIT ITERATIVE TECHNIQUES*

JAMES E. GUNN†

1. Introduction. In [8], the author discusses an iterative scheme for solving a difference analogue for the elliptic differential equation $\nabla \cdot a \nabla u = f$ on two-dimensional rectangular regions with Dirichlet boundary conditions. It is shown there that a semi-explicit technique involving the inversion only of the Peaceman-Rachford [10] alternating-direction operators for the Laplacian gives convergence in $O(h^{-2} \log h^{-1})$ operations.

The results of that paper are here extended to a general class of semi-explicit iterative techniques for not-necessarily-symmetric operators and application is made to the difference analogues of differential equations of the form

$$\sum_{i=1}^{m} \frac{\partial}{\partial x_i}\left(a_i \frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{m} b_i \frac{\partial u}{\partial x_i} + f(x_1, \cdots, x_m, u) = 0,$$

on rectangular regions in Cartesian m-space. A computing estimate of $O(h^{-m} \log h^{-1})$ is again obtained, with an explicit estimate in terms of the number of operations required to solve the m-dimensional Poisson's equation by generalized alternating directions [5].

For linear equations with self-adjoint operators, a Čebyšev iteration is presented for which one obtains a computing estimate considerably better than that of [8].

Finally, the results of numerical experiments are discussed for the equation $\nabla \cdot a \nabla u = f$ on a cube, and comparisons are made with the Douglas-Brian alternating-direction method [2], [4] and the method of successive overrelaxation [11].

2. The semi-explicit iterative technique. Let us consider obtaining the solution of

$$Dx = y, \tag{2.1}$$

for x, y belonging to a finite-dimensional complex inner product space $\mathfrak{C}$ (hereafter referred to as a *unitary space*), and D a nonsingular operator on $\mathfrak{C}$. We require that

$$\operatorname{Re} D = \tfrac{1}{2}(D + D^*) = B > 0, \tag{2.2}$$

i.e., $(u, Bu) > 0$ for all $u \neq 0$ in $\mathfrak{C}$.

Let A be a positive-definite operator on $\mathfrak{C}$ whose inverse is known (or can be approximated in a self-adjoint manner—we shall discuss this later).

* Received by the editors January 6, 1964, and in revised form August 24, 1964.

† Humble Oil and Refining Company. Now at 208 Robinson Laboratory, California Institute of Technology, Pasadena, California.

Since $\mathfrak{H}$ is unitary, D and A are bounded and compact, and the compactness of the unit sphere in $\mathfrak{H}$ implies the existence of positive constants $\varkappa_1$, $\varkappa_2$, and $\varkappa_3$ such that

$$\begin{aligned} \varkappa_1(u, Au) &\leqq (u, Bu) \leqq \varkappa_2(u, Au), \\ |(u, Cu)| &\leqq \varkappa_3(u, Au), \end{aligned} \tag{2.3}$$

for all nonzero u in $\mathfrak{H}$. Here $C = \operatorname{Im} D = (1/2i)(D - D^*)$. It is to one's advantage to choose A in such a fashion that $\varkappa_1$ is close to $\varkappa_2$ and $\varkappa_3$ is small, as will be evident in the analysis to follow. We define the semi-explicit iterative scheme ([3], [8]) as follows:

$$Ax^{n+1} = Ax^n - \rho(Dx^n - y), \tag{2.4}$$

where ρ is a positive iteration parameter to be chosen later. The solution x of (2.1) is clearly a fixed point of (2.4), and the error e^n obeys

$$Ae^{n+1} = (A - \rho D)e^n.$$

We must show that we can choose ρ so that the e^n go to zero in some norm. We shall treat nonlinear problems later, in which one encounters equations like (2.4) with a sequence $\{D_n\}$; we thus prove a slightly more powerful result than needed:

THEOREM 1. *Let $\{D_n\}$ be a sequence of nonsingular operators on $\mathfrak{H}$ satisfying (2.2) for each n. Let A be positive-definite and satisfy (2.3) with $D = D_n$, $\varkappa_1$, $\varkappa_2$, $\varkappa_3$ fixed. Then there exists a positive ρ such that the sequence $\{u^n\}$ defined by*

$$Au^{n+1} = (A - \rho D_n)u^n \tag{2.5}$$

tends to zero in the A-norm $\| u^n \|_A^2 = (u^n, Au^n)$, and such that the number of iterations necessary to reduce the A-norm of the initial error by a factor ϵ is

$$O\left(\frac{4\varkappa_3^2 + \varkappa_1(\varkappa_1 + \varkappa_2)}{2\varkappa_1^2}\right) \log \epsilon^{-1}.$$

We shall need the following lemmas.

LEMMA 1. *Let N be a normal operator on a unitary space $\mathfrak{U}$. Let $R = \frac{1}{2}(N + N^*)$, $K = (1/2i)(N - N^*)$. Then for any $u \in \mathfrak{U}$, $|(u, (R + iK)u| = \sqrt{(u, Ru)^2 + (u, Ku)^2}$; furthermore, $\| R + iK \|^2 \leqq \| R^2 \| + \| K \|^2$.*

Proof. The proof is simple, depending only upon the fact that a complex number is equal to the square root of the sum of the squares of its real and imaginary parts, and will be omitted.

LEMMA 2. *Let A be a positive and R a self-adjoint operator on $\mathfrak{U}$, satisfying $\alpha(u, Au) \leqq |(u, Ru)| \leqq \beta(u, Au)$. Then $\| A^{-1/2}RA^{-1/2} \| \leqq \max(|\alpha|, |\beta|)$.*

Proof. $A^{-1/2}RA^{-1/2}$ is itself a self-adjoint operator, so there exists a com-

plete orthonormal set $\{\psi_i\}$ in $\mathcal{U}$ such that $A^{-1/2}RA^{-1/2}\psi_i = \lambda_i\psi_i$, λ_i real. Then $\| A^{-1/2}RA^{-1/2} \| = \max_i | \lambda_i |$; but if $\varphi_i = A^{-1/2}\psi_i$, then

$$(\psi_i, A^{-1/2}RA^{-1/2}\psi_i) = (\varphi_i, R\varphi_i) = \lambda_i(\varphi_i, A\varphi_i).$$

Since $(\varphi_i, A\varphi_i) > 0$, we can divide and obtain

$$\max_i | \lambda_i | = \max_i \frac{(\varphi_i, R\varphi_i)}{(\varphi_i, A\varphi_i)} \leqq \max (|\alpha|, |\beta|).$$

The proof of the theorem now follows. If we set $v^n = A^{1/2}u^n$, then (2.5) becomes

$$v^{n+1} = v^n - \rho A_n^{-1/2}D_nA_n^{-1/2}v^n = Q_nv^n. \tag{2.6}$$

Let $B_n = \frac{1}{2}(D_n + D_n^*)$, $C_n = \frac{1}{2}(D_n - D_n^*)$. Then

$$v^{n+1} = (1 - \rho A^{-1/2}(B_n + C_n)A^{-1/2})v^n, \tag{2.7}$$

and B_n and C_n satisfy (2.3) for each n. The triangle inequality yields

$$\begin{aligned} \| 1 - \rho A^{-1/2}(B_n + C_n)A^{-1/2} \| \leqq \|(1 - \theta) - \rho A^{-1/2}B_nA^{-1/2} \| \\ + \| \theta - \rho A^{-1/2}C_nA^{-1/2} \| \end{aligned} \tag{2.8}$$

for all real numbers θ. By Lemmas 1 and 2, we have, for θ positive and less than 1,

$$\| \theta - \rho A^{-1/2}C_n A^{-1/2} \| \leqq \theta \sqrt{1 + \rho^2\left(\frac{\kappa_3}{\theta}\right)^2} \leqq \theta + \frac{\rho^2\kappa_3^2}{2\theta}. \tag{2.9}$$

Applying Lemma 2 once more, we have

$$\begin{aligned} \|(1 - \theta) - \rho A^{-1/2}B_nA^{-1/2} \| \leqq \max (|1 - \theta - \rho\kappa_1|, \\ |1 - \theta - \rho\kappa_2|). \end{aligned} \tag{2.10}$$

Thus

$$\begin{aligned} &\| 1 - \rho A^{-1/2}(B_n + C_n)A^{-1/2} \| \\ &\quad \leqq \min_{0<\theta<1} \left\{ \theta + \frac{\rho^2\kappa_3^2}{2\theta} + \max (|1 - \theta - \rho\kappa_1|, |1 - \theta - \rho\kappa_2|) \right\}. \end{aligned} \tag{2.11}$$

It is clear that by taking ρ sufficiently small, this becomes $1 - \rho\kappa_1 + \rho^2\kappa_3^2/2\theta$ which is less than one for any $\theta > 0$ for sufficiently small ρ. We can do a little better than this, however, and find the ρ, θ combination which minimizes the right-hand side of (2.11). Let

$$f(\theta, \rho) = \theta + \frac{\rho^2\kappa_3^2}{2\theta} + \max \{|1 - \theta - \rho\kappa_1|, |1 - \theta - \rho\kappa_2|\}. \tag{2.12}$$

We wish to minimize f in the strip $0 \leqq \theta \leqq 1$, $\rho \geqq 0$. We must distinguish two regions, one (I) where $\rho\varkappa_2 + \theta - 1 \leqq 1 - \theta - \rho\varkappa_1$ and the other (II) where $\rho\varkappa_2 + \theta - 1 > 1 - \theta - \rho\varkappa_1$. We find easily that (I) is a triangle bounded by the ρ axis, the θ axis, and the line $\rho = (2 - 2\theta)/(\varkappa_1 + \varkappa_2)$. In (I),

$$f(\theta, \rho) = 1 - \rho\varkappa_1 + \frac{\rho^2\varkappa_3^2}{2\theta},$$

which is a decreasing function of θ for each ρ. Thus the minimum of the function in this region must occur on the right boundary, the line $\rho = (2 - 2\theta)/(\varkappa_1 + \varkappa_2)$. In (II),

$$f(\theta, \rho) = \rho\varkappa_2 + 2\theta - 1 + \frac{\rho^2\varkappa_3^2}{2\theta},$$

which is an increasing function of ρ for each θ, so the minimum here must occur on the lower boundary, again the line $\rho = (2 - 2\theta)/(\varkappa_1 + \varkappa_2)$. Thus we must minimize $f(\theta, (2 - 2\theta)/(\varkappa_1 + \varkappa_2))$ as a function of θ. This is trivially done, and we find

$$\theta^2 = \frac{\varkappa_3^2}{\varkappa_3^2 + \varkappa_1(\varkappa_1 + \varkappa_2)}, \qquad \rho = \frac{2}{\varkappa_1 + \varkappa_2}\left(1 - \sqrt{\frac{\varkappa_3^2}{\varkappa_3^2 + \varkappa_1(\varkappa_1 + \varkappa_2)}}\right),$$

and

$$\| 1 - \rho A^{-1/2}(B_n + C_n)A^{-1/2} \| \leqq 1 - \frac{2\varkappa_3^2}{(\varkappa_1 + \varkappa_2)^2}\left(\sqrt{1 + \varkappa_1(\varkappa_1 + \varkappa_2)/\varkappa_3^2} - 1\right)^2 < 1, \tag{2.13}$$

which holds also in the case $\varkappa_3 = 0$ (D_n self-adjoint), in the form

$$\| 1 - \rho A^{-1/2}(B_n + C_n)A^{-1/2} \| \leqq \frac{\varkappa_2 - \varkappa_1}{\varkappa_2 + \varkappa_1}.$$

Thus we have

$$\| v^n \| \leqq \left[1 - \frac{2\varkappa_3^2}{(\varkappa_1 + \varkappa_2)^2} \cdot \left(\sqrt{1 + \frac{\varkappa_1(\varkappa_1 + \varkappa_2)}{\varkappa_3^2}} - 1\right)^2\right]^n \| v^0 \| \to 0 \quad \text{as} \quad n \to \infty,$$

but $\| v^n \| = \| A^{1/2}u^n \| = (u^n, Au^n)^{1/2}$. To reduce the A-norm of v below ϵ it is necessary only to take n such that

$$\left[1 - \frac{2\varkappa_3^2}{(\varkappa_1 + \varkappa_2)^2}\left(\sqrt{1 + \frac{\varkappa_1(\varkappa_1 + \varkappa_2)}{\varkappa_3^2}} - 1\right)^2\right]^n < \epsilon, \tag{2.14}$$

or

$$n \geqq \frac{\log \epsilon}{\log\left[1 - \frac{2\kappa_3^2}{(\kappa_1+\kappa_2)^2}\left(\sqrt{1 - \frac{(\kappa_1+\kappa_2)\kappa_1}{\kappa_3^2}} - 1\right)^2\right]} \tag{2.15}$$

$$\sim \frac{(\kappa_1+\kappa_2)^2}{2\kappa_3^2}\left(\sqrt{1 + \frac{(\kappa_1+\kappa_2)\kappa_1}{\kappa_3^2}} - 1\right)^{-2} \log \epsilon^{-1},$$

as $\kappa_1 \to 0$ or κ_2 or $\kappa_3 \to \infty$; the estimate is quite good enough for computational purposes whenever $\min(2\kappa_1/\kappa_2, \kappa_1^2/4\kappa_3^2)$ is less than $\frac{1}{2}$ or so, and in turn is estimated to within 30 per cent by

$$n \sim \frac{4\kappa_3^2 + \kappa_1(\kappa_1+\kappa_2)}{2\kappa_1^2} \log \epsilon^{-1}.$$

(This corresponds to the approximation $(2/x)(2 + x - 2\sqrt{1+x}) \doteqdot 2x/(x+4)$, which maintains the stated accuracy for all $x > 0$.) This completes the proof of the theorem.

COROLLARY. *If*

$$\rho = \frac{2}{\kappa_1+\kappa_2}\left(1 - \sqrt{\frac{\kappa_3^2}{\kappa_3^2 + \kappa_1(\kappa_1+\kappa_2)}}\right),$$

the scheme (2.4) *converges with the rate* (2.15) *in the A-norm to the solution* x *of* (2.1).

It is clear that we need something more than Theorem 1 to treat most nonlinear problems, since if we wish to solve $D(x) = f$ by the scheme (2.4), the error $e^n = x^n - x$ satisfies

$$Ae^{n+1} = Ae^n - \rho(D(x^n) - D(x)).$$

For Theorem 1 to be applicable, it is evident that some Lipschitz condition must exist for $D(u)$, and it must in general be in some sense "almost linear." For example, if we can write $D(x^n) - D(x) = \hat{D}(\bar{x}^n)e^n$ by some variant of the mean-value theorem, and obtain estimates like (2.3) uniformly for $\hat{D}(x)$, the procedure converges. It is precisely this situation we shall treat in §3.

It will often happen that the obvious choice for A is not itself easily invertible, but if A can be expressed as the sum of commuting operators A_i, each positive semi-definite with the sum positive on $\mathfrak{H}$, we can approximate A by an inner "alternating-direction" iteration [5]. The operators used to approximate A are functions of the A_i and are hence self-adjoint and commutative with A. Let us consider such a scheme or a related one. We wish to solve

$$Af = g. \tag{2.16}$$

Let Λ be the error propagator for an iterative scheme for the solution of (2.16); i.e., if f^n is the nth iterate,

$$(f^{n+1} - f) = \Lambda(f^n - f). \tag{2.17}$$

We need assume only that Λ is self-adjoint and commutes with A. In this case we clearly have

$$Af^{n+1} = \Lambda Af^n + (1 - \Lambda)g. \tag{2.18}$$

It is clear that the operator Λ can represent one or more ordinary iteration steps; in particular, it can represent a cycle of alternating-direction iteration using a sequence of iteration parameters. If we apply one step of the above scheme, which we shall designate by (Λ), to obtain an estimate for x^{n+1} in (2.4), we obtain the new scheme

$$Ax^{n+1} = Ax^n - \rho(1 - \Lambda)(Dx^n - y), \tag{2.19}$$

or, in more tractable form,

$$(1 - \Lambda)^{-1}Ax^{n+1} = (1 - \Lambda)^{-1}Ax^n - \rho(Dx^n - y). \tag{2.20}$$

Now in order for scheme (Λ) to converge, we must have the spectral radius of Λ less than 1, but since Λ is self-adjoint, this means that $\| \Lambda \| = \zeta < 1$, and $(1 - \Lambda)^{-1}$ is positive-definite. Since $(1 - \Lambda)$ commutes with A, $(1 - \Lambda)^{-1}A$ is also positive-definite, and the analysis of this scheme reduces to that of (2.4), if we can show that (2.3) holds for the operator $(1 - \Lambda)^{-1}A$. This is not difficult to see, however, since

$$\begin{aligned} (u, Au) &= (u, (1 - \Lambda)(1 - \Lambda)^{-1}Au) \\ &= ((1 - \Lambda)^{-1/2}A^{1/2}u, (1 - \Lambda)(1 - \Lambda)^{-1/2}A^{1/2}u) \\ &\leqq \| 1 - \Lambda \| \, \|(1 - \Lambda)^{-1/2}A^{1/2}u \|^2 \\ &\leqq (1 + \zeta)(u, (1 - \Lambda)^{-1}Au) \end{aligned} \tag{2.21}$$

and, similarly,

$$\begin{aligned} (u, (1 - \Lambda)^{-1}Au) &\leqq \| (1 - \Lambda)^{-1} \| (u, Au) \\ &\leqq \frac{1}{1 - \zeta} (u, Au). \end{aligned} \tag{2.22}$$

Thus (2.3) holds with $\varkappa_1$, $\varkappa_2$, $\varkappa_3$ replaced by $(1 - \zeta)\varkappa_1$, $(1 + \zeta)\varkappa_2$, $(1 + \zeta)\varkappa_3$, respectively. We can estimate the number of iterations necessary to solve (2.1) by (2.19) in terms of the number required to solve (2.16) by (2.18). The equation obeyed by the error $e_n = x^n - x$ for (2.19) is clearly

$$A(1 - \Lambda)^{-1}e^{n+1} = \{A(1 - \Lambda)^{-1} - \rho D\}e^n, \tag{2.23}$$

and that for $\epsilon^n = f^n - f$ for (2.18) is (2.17). Note first that $\eta^n = A^{1/2}\epsilon^n$ satisfies

$$\eta^{n+1} = A^{1/2}\Lambda A^{-1/2}\eta^n = \Lambda\eta^n, \tag{2.24}$$

and $\| \eta^n \| = \| \epsilon^n \|_A$. Thus, the number of iterations necessary to reduce the A-norm of the initial error in (2.24) by ϵ can be approximated by

$$N_\Lambda \doteqdot \frac{\log \epsilon}{\log \zeta},$$

and the number N_D required to do the same for (2.23) is

$$N_D \doteqdot \frac{\log \epsilon}{\log\left[1 - \dfrac{2(1+\zeta)^2\varkappa_3^2}{[(1-\zeta)\varkappa_1 + (1+\zeta)\varkappa_2]^2} \cdot \left(\sqrt{1 + \dfrac{\{(1-\zeta)\varkappa_1 + (1+\zeta)\varkappa_2\}(1-\zeta)\varkappa_1}{(1+\zeta)^2\varkappa_3^2}} - 1\right)^2\right]}$$

$$\doteqdot \frac{4(1+\zeta)^2\varkappa_3^2 + \varkappa_1(1-\zeta)\{\varkappa_1(1-\zeta) + \varkappa_2(1+\zeta)\}}{2(1-\zeta)^2\varkappa_1^2} \log \epsilon^{-1}.$$

If ζ is not too much different from 1, the ratio N_D/N_Λ becomes approximately

$$\frac{16\varkappa_3^2 + (1-\zeta^2)\varkappa_1\varkappa_2}{2(1-\zeta)\varkappa_1^2}. \tag{2.25}$$

In the important case when $\varkappa_3 = 0$,

$$\frac{N_D}{N_\Lambda} \doteqdot \frac{\varkappa_2}{\varkappa_1}. \tag{2.26}$$

3. Applications to elliptic partial difference equations.

(i) *The Dirichlet problem for mildly nonlinear equations.* Let us consider the Dirichlet problem for

$$\sum_{i=1}^{m} -\frac{\partial}{\partial x_i}\left(a_i(\bar{x})\frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{m} b_i(\bar{x})\frac{\partial u}{\partial x_i} + f(\bar{x}, u) = 0 \tag{3.1}$$

on a rectangular region R in Cartesian m-space; here $\bar{x} = (x_1, \cdots, x_m)$, and $u = g(\bar{x})$ on ∂R. We approximate (3.1) by the usual five-point difference system on a rectangular net R_h on R, using the following difference operators:

$$\begin{aligned} \nabla_i u &= \frac{u(x_1, \cdots, x_i + h_i, \cdots, x_m) - u(x_1, \cdots, x_i, \cdots, x_m)}{h_i}, \\ \bar{\nabla}_i u &= \nabla_i u(x_1, \cdots, x_i - h_i, \cdots, x_m), \\ \hat{\nabla}_i u &= \tfrac{1}{2}(\nabla_i + \bar{\nabla}_i)u, \end{aligned} \tag{3.2}$$

where h_i is the mesh size in the i-direction. With this notation, the difference analogue of (3.1) becomes

$$(3.3)\quad L_h(u) = \sum_{i=1}^{m} -\bar{\nabla}_i(\hat{a}_i(\bar{x})\nabla_i u(\bar{x})) + \sum_{i=1}^{m} b_i(\bar{x})\hat{\nabla}_i u(\bar{x}) + \hat{f}(\bar{x}, u) = 0,$$

where $\hat{a}_i(\bar{x}) = a_i(x_i, \cdots, x_i + h_i/2, \cdots, x_m)$, and $\hat{f}(\bar{x}, u)$ contains an additional inhomogeneous term at points adjacent to the boundary representing the boundary condition $u = g(\bar{x})$. The boundary conditions are also reflected in the definitions of the operators in (3.3); the boundary terms are simply missing, since they become part of the inhomogeneous term.

We define the inner product

$$(3.4)\quad (\varphi, \psi) = \prod_{k=1}^{m} h_k \sum_{R_h} \varphi(\bar{x})\overline{\psi(\bar{x})}.$$

With this inner product, the space $\mathfrak{C}$ of all complex-valued functions defined on the net R_h is clearly a unitary space of dimension equal to the number of points N of the net. Then the operators in (3.2) become $N \times N$ matrices. We shall use the same symbols for the difference operators, whose operands are values of net functions, and the corresponding matrix operator, whose operand is a member of $\mathfrak{C}$; whenever confusion is likely to arise, we shall use the argument $\bar{x}$ explicitly when speaking of the former.

Note that the definition of the difference operators at the points adjacent to the boundary is such that $-\bar{\nabla}_i\hat{a}\nabla_i$ is *not* the matrix product of $\bar{\nabla}_i$, the diagonal matrix $\hat{a}_i$ and ∇_i; one verifies easily, however, that $-\bar{\nabla}_i\hat{a}_i\nabla_i$ is a positive-definite matrix (see §3(iii) for the computation; here the θ in that discussion is zero) for $\hat{a}(\bar{x})$ an everywhere positive function. Under the conditions that $a(\bar{x})$ be positive, $\partial f/\partial u$ exist and be positive, Bers [1] has shown that the solution of (3.3) is unique and converges to the solution of (3.1) as $\max_i h_i \to 0$. We concern ourselves with the algebraic problem of obtaining a solution of (3.3). We assume that each of the functions f, a_i, and b_i appearing in (3.1) is real and that

$$(3.5)\quad \begin{aligned} &\text{(a)} && 0 < \mu_{0i} \leqq a_i(x) \leqq \mu_i^0 \quad \text{in} \quad R_h, \\ &\text{(b)} && \frac{\mu_i^0}{\mu_{0i}} \leqq M_1, \\ &\text{(c)} && 0 \leqq \frac{\partial f}{\partial u} \leqq M_2 \quad \text{in} \quad R_h, \\ &\text{(d)} && \left|\frac{\partial b_i}{\partial x_i}\right| \leqq 2\eta\mu_{0i}\lambda_i, \qquad 0 \leqq \eta < 1, \\ &\text{(e)} && |b_i| < M_4. \end{aligned}$$

Condition (c) is rather stronger than needed, but simplifies the analysis somewhat. The λ_i appearing in (d) are the minimal eigenvalues of $-\overline{\nabla}_i\nabla_i$, well known to be $(4/h_i^2)\sin^2(\pi h_i/2l_i)$, where l_i is the length of the region in the i-direction. This condition arises naturally from our discussion but is not necessary for the existence of u and it is possible that it could be removed by a more careful analysis. Let us define the iterator A by

$$Au = -\sum_{i=1}^{m} (\mu_i{}^0\overline{\nabla}_i\nabla_i u) + \gamma u, \tag{3.6}$$

where γ is a constant that will be specified later.

It is well known that the operators $-\overline{\nabla}_i\nabla_i$ commute among themselves and are positive-definite on rectangular regions; the inversion of one such operator requires the solution of a tridiagonal matrix equation and so is quite easily done. We may thus define an alternating-direction iteration [5] for the solution of

$$Au = y, \tag{3.7}$$

possessing an error propagator Λ for a cycle of variable parameters which is self-adjoint, commutes with A, and has a norm in the vicinity of $\frac{1}{2}$ for a number of iterations per cycle about half the logarithm of the ratio of the maximum to the minimum eigenvalues of A. (For a given region, this ratio is $O([\min_i h_i]^{-2})$.) The semi-explicit iteration scheme for (3.3) thus becomes

$$(1 - \Lambda)^{-1}Au^{n+1} = (1 - \Lambda)^{-1}Au^n - \rho L_h(u^n), \tag{3.8}$$

where u^{n+1} is obtained from u^n by performing one cycle of alternating-direction iteration on (3.7), with the initial estimate u^n and $y = Au^n - \rho L_h(u^n)$. It is clear that the error $e^n = u^n - u$ satisfies

$$\begin{aligned}(1 - \Lambda)^{-1}Ae^{n+1} \\ = (1 - \Lambda)^{-1}Ae^n - \rho\left\{\sum_{i=1}^{m} - \overline{\nabla}_i\,\hat{a}_i\,\nabla_i\,e^n + \sum_{i=1}^{m} b_i\,\hat{\overline{\nabla}}_i\,e^n + f_u{}^{(n)}e^n\right\},\end{aligned} \tag{3.9}$$

where $f_u{}^{(n)}$ depends on u and u_n and has a value at $\bar{x}$ between $(\partial f/\partial u)(u, \bar{x})$ and $(\partial f/\partial u)(u^n, \bar{x})$. Let $D_n e^n$ be the expression in braces.

To demonstrate the convergence of (3.8) it is clearly necessary only to show that D_n has a positive-definite real part; to obtain the rate we will need estimates for the $\varkappa$'s in (2.3).

The only troublesome term in D_n is the first-order one; it is skew for $b_i(\bar{x})$ constant on R_h, but if b_i is variable it acquires a real part proportional to $\partial b_i/\partial x_i$. A straightforward computation yields the result that, for $u \in \mathfrak{H}$,

$$(u, \{\mathrm{Re}\, b_i \hat{\nabla}_i\}u) = \frac{1}{2}(u, b_i \hat{\nabla}_i u - \hat{\nabla}_i b_i u) \leqq \frac{1}{2}\sup_R \left|\frac{\partial b_i}{\partial x_i}\right| \cdot \| u \|^2,$$

$$(3.10)\quad (u, \{\mathrm{Im}\, b_i \hat{\nabla}_i\}u) = \frac{1}{2}(u, b_i \hat{\nabla}_i u + \hat{\nabla}_i b_i u)$$

$$\leqq \frac{1}{2}\sup_R \left|\frac{\partial b_i}{\partial x_i}\right| \cdot \| u \|^2 + \sup | b_i | \, \| \hat{\nabla}_i u \| \cdot \| u \|.$$

One also finds quite easily that

$$(3.11)\qquad \| \hat{\nabla}_i u \| \leqq (u, -\overline{\nabla}_i \nabla_i u)^{1/2},$$

and it follows from the minimal property of the least eigenvalue that

$$(3.12)\qquad \| u \|^2 \leqq \frac{1}{\lambda_i}(u, -\overline{\nabla}_i \nabla_i u).$$

Thus, using (3.5d),

$$(u, \{\mathrm{Re} \sum_i b_i \hat{\nabla}_i\}u) \leqq \frac{1}{2}\sum_i \frac{1}{\lambda_i}(u, -\overline{\nabla}_i \nabla_i u) \sup_R \left|\frac{\partial b_i}{\partial x}\right|$$

$$(3.13)\qquad \leqq \eta \sum_i \mu_{0i}(u, -\overline{\nabla}_i \nabla_i u)$$

$$\leqq \eta \sum_i (u, -\overline{\nabla}_i a_i \nabla_i u).$$

The last inequality can be established easily, as can a similar one for an upper bound,

$$(3.14)\qquad \sum_i (u, \overline{\nabla}_i \hat{a}_i \nabla_i u) \leqq \sum \mu_i{}^0(u, -\overline{\nabla}_i \nabla_i u),$$

by expanding the inner product $(u, \overline{\nabla}_i \hat{a}_i \nabla_i u)$. (See §3(iii).) Using (3.5), (3.13), and (3.14), we see finally that

$$(3.15)\quad (u, \{\mathrm{Re}\, D_n\}u) \leqq (1 + \eta) \sum_i \mu_i{}^0(u, -\overline{\nabla}_i \nabla_i u) + M_2(u, u);$$

$$(u, \{\mathrm{Re}\, D_n\}u) \geqq (1 - \eta) \sum_i \mu_{i0}(u, -\overline{\nabla}_i \nabla_i u)$$

$$(3.16)$$

$$\geqq \frac{(1 - \eta)}{M_1} \sum_i \mu_i{}^0(u, -\overline{\nabla}_i \nabla_i u).$$

Let γ in (3.6) be equal to $M_2/2(1 + \eta)$. Using (3.12), we obtain

$$(3.17)\qquad (u, u) \leqq \frac{\sum_i \mu_i{}^0(u, -\overline{\nabla}_i \nabla_i u)}{\sum_i \lambda_i \mu_i{}^0}.$$

Then, setting $\lambda = \sum_{i=1}^m \mu_i{}^0 \lambda_i$,

$$\frac{1-\eta}{M_1\left(1+\frac{\gamma}{\lambda}\right)}(u, Au) \leqq (u, \{\mathrm{Re}\, D_n\}u) \leqq (1+\eta)\left(1+\frac{\gamma}{\lambda}\right)(u, Au). \tag{3.18}$$

Hence we can set

$$\varkappa_1 = \frac{1-\eta}{M_1\left(1+\frac{\gamma}{\lambda}\right)}, \qquad \varkappa_2 = (1+\eta)\left(1+\frac{\gamma}{\lambda}\right). \tag{3.19}$$

In a similar manner, we can show that

$$\varkappa_3 = \eta + M_4 \sum_{i=1}^{m} \frac{1}{\mu_i^0 \sqrt{\lambda_i}} \tag{3.20}$$

is a bound for the skew part of D_n in terms of A. We note immediately that each of these numbers is independent, or nearly so, of the mesh size h_i—the mesh size enters only through the λ_i, which are very nearly $(\pi/l_i)^2$ for all small h_i. Thus the ratio (2.25) of the number of tridiagonal inversions necessary to solve (3.3) by the semi-explicit method to the number necessary to solve (3.7) by alternating directions is very nearly independent of h, and becomes independent of h_i as $\max_i h_i$ tends to zero. Since the number of operations involved in these inversions necessary to reduce the norm of the initial error in (3.7) by a factor ϵ is known to be

$$O\left(\left\{\prod_{i=1}^{m} h_i\right\} \log\,(\min_i h_i)\, \log \epsilon\right),$$

the number required for (3.3) is no larger than a constant, independent of h, times this. In the case where $b_i = 0$, $\eta = 0$, this constant may easily be seen to be in the vicinity of M_1, the maximum of the ratio of the maximum of a_i to the minimum of a_i in R. A number of experiments have indicated that this estimate is very conservative, however; we shall discuss this later.

(ii) *Linear equations with no skew term.* A decided improvement on the rate obtained in the last section is possible for elliptic difference equations of the form

$$L_h u = \sum_{i=1}^{m} -\bar{\nabla}_i \hat{a}_i \nabla_i u + qu = f, \tag{3.21}$$

where $0 \leqq q(x) \leqq M_2$ and is independent of u, as is f. If we set up the iterative scheme

$$(1-\Lambda)^{-1} A u^{n+1} = (1-\Lambda)^{-1} A u^n - \rho_n (L_h u^n - f), \tag{3.22}$$

where A is defined as in (3.6) with $\gamma = M_2/2$, and set $e_n = A^{1/2}(1 - \Lambda)^{-1/2} \cdot (u^n - u)$, then

$$e^{n+1} = (1 - \rho_n A^{-1/2}(1 - \Lambda)^{-1/2} L_h (1 - \Lambda)^{-1/2} A^{-1/2}) e^n. \tag{3.23}$$

The operator appearing in the second term on the right is positive-definite, and so has a complete set of eigenvectors φ_j such that

$$A^{-1/2}(1 - \Lambda)^{1/2} L_h (1 - \Lambda)^{1/2} A^{-1/2} \varphi_j = \nu_j \varphi_j$$

and

$$(\varphi_j , \varphi_k) = \delta_{jk} .$$

It is clear from Lemma 2 and the results of the preceding section that the ν_j are bounded above and below by

$$(1 + \zeta)\left(1 + \frac{M_2}{2\lambda}\right) \quad \text{and} \quad (1 - \zeta)\left(1 + \frac{M_2}{2\lambda}\right)^{-1} M_1^{-1},$$

respectively. (Recall that $\zeta = \| \Lambda \|$, $M_1 = \max_i (\mu_0{}^i/\mu_{0i})$.) We can express the e^n as linear combinations of the φ_i,

$$e^n = \sum_j d_j{}^n \varphi_j ,$$

and obtain

$$d_j{}^{n+1} = (1 - \rho_n \nu_j)\, d_j{}^n. \tag{3.24}$$

Since $\| e^n \|^2 = \sum (d_j{}^n)^2$,

$$\| e^p \| \leqq \max_j \prod_{n=0}^{p-1} (1 - \rho_n \nu_j) \| e^0 \|. \tag{3.25}$$

The problem of choosing the sequence ρ_n so as to minimize the maximum of the polynomial $\prod_{n=0}^{p-1} (1 - \rho_n \nu)$ for ν in the interval $[a, b]$ is a well-known one and has a well-known solution [9], [12]; the minimum of the maximum occurs when the ρ_n are chosen to be the roots of the pth order Čebyšev polynomial $T_p(\nu') = \cos (p \cos^{-1} \nu')$ in the variable

$$\nu' = \frac{2\nu}{b - a} - \frac{b + a}{b - a} .$$

Explicitly,

$$\rho_n = 2\left[(b + a) - (b - a) \cos \frac{(2n - 1)\pi}{2p}\right]^{-1}. \tag{3.26}$$

The maximum of the product is then given by

$$\max_{\nu \in [a,b]} \prod_{n=0}^{p-1} (1 - \rho_n \nu) = \left[T_p\left(\frac{b+a}{b-a}\right)\right]^{-1} = \left[\cosh\left(p \cosh^{-1}\left(\frac{b+a}{b-a}\right)\right)\right]^{-1}. \tag{3.27}$$

If

$$\frac{b}{a} = \frac{(1+\zeta)\left(1+\frac{M_2}{2\lambda}\right)^2 M_1}{1-\zeta}$$

is large, then choosing p in the vicinity of $(b/a)^{1/2}$ and using this sequence of ρ_n's cyclically requires only approximately $\frac{1}{2}\sqrt{b/a}\log\epsilon$ iterations as opposed to $\frac{1}{2}(b/a)$ iterations for an "optimum" fixed ρ to reduce the A-norm of the error by a factor ϵ. This estimate is a simple consequence of the expression (3.27), and the analysis leading to it will be omitted; a similar situation is discussed in [12], and proofs and references for the minimization problem are given there.

The scheme outlined above bears a strong superficial relationship to the Young-Richardson relaxation method [12], but is of course much faster; the problems encountered there with instability against round-off also occur here. No experiments have as yet been conducted to determine the best order for the ρ_n's, but it is likely that the order proposed by Young in the above reference will suffice. There one starts in the middle of the range of ρ's and works up and down on alternate steps, terminating with the largest.

(iii) *The Neumann and Robin problems.* The problem for more complicated boundary conditions than Dirichlet is in most respects quite similar to the Dirichlet case; the difference equation (3.3) is unchanged in form, though the definition of the difference operators is changed at points adjacent to the boundary. The Neumann problem presents a difficulty of its own, since the second-order difference operators for this problem are singular and hence are no longer positive-definite. This we shall consider later; but we look for the present at the general (Robin) boundary-value problem, where one knows

$$\alpha(\bar{x})u + \beta(\bar{x})\frac{\partial u}{\partial n} = g(\bar{x}) \tag{3.28}$$

on ∂R. We shall consider the form of the operators ∇ and $\bar{\nabla}$ in the one-dimensional case; the generalization to m dimensions is immediate. We again construct a grid of size h upon the interval R, but now, if $R = [a, b]$, we place the grid points at $x_1 = a + h/2$, $x_2 = a + 3h/2$, $\cdots$, $x_k = a + (2k - 1)h/2$, with $x_N = b - h/2$. This is a standard choice for the Neumann problem; it allows us to approximate $\partial u/\partial n$ by $[u(x_0) - u(x_1)]/h$

with an error which is $O(h^2)$. One can clearly also use this grid for Dirichlet boundary conditions, specifying $[u(x_0) + u(x_1)]/2$ as the boundary value, again with an $O(h^2)$ error. It is now clear how one represents (3.28) in the discretized case; we specify

$$u(x_0) + \theta u(x_1) = \hat{g}, \tag{3.29}$$

where

$$\theta = \frac{\alpha/2 - \beta/h}{\alpha/2 + \beta/h}, \qquad \hat{g}(x) = \frac{g}{\alpha/2 + \beta/h}.$$

Here we require that $\alpha\beta \geqq 0$, a condition known to be necessary for stability in the differential case; thus, $-1 \leqq \theta \leqq 1$. Since the combination $u_0 + \theta(x)u_1$ is known, the operator $\bar{\nabla}$ at the point x_1, the first interior point, is defined as

$$\bar{\nabla}u_1 = \frac{(1 + \theta)u_1}{h}, \tag{3.30}$$

with an analogous expression for ∇u_N :

$$\nabla u_N = \frac{(1 + \theta)u_N}{h}. \tag{3.31}$$

The operator $\bar{\nabla}\hat{a}\nabla$ is, as before, not the matrix product of $\bar{\nabla}$, $\hat{a}$, and ∇, but is still symmetric and nonnegative-definite for $-1 \leqq \theta \leqq 1$; its definition for u_1 is

$$\nabla\hat{a}\nabla u_1 = \frac{1}{h^2}\{\hat{a}_1(u_2 - u_1) - \hat{a}_0(1 + \theta)u_1\}, \tag{3.32}$$

with an analogous expression for u_N . The inner product $(u, \bar{\nabla}\hat{a}\nabla u)$ becomes

$$\begin{aligned}(u, -\bar{\nabla}\hat{a}\nabla u) = {} & -u_1\overline{\{\hat{a}_1(u_2 - u_1) - \hat{a}_0(1+ \theta)u_1\}} \\ & - u_2\overline{\{\hat{a}_2(u_3 - u_2) - \hat{a}_1(u_2 - u_1)\}} \\ & \cdots - u_N\overline{\{-\hat{a}_N(1 + \theta)u_N - \hat{a}_{N-1}(u_N - u_{N-1})\}}.\end{aligned}$$

Summing by parts, we obtain

$$\begin{aligned}(u, -\bar{\nabla}a\nabla u) = {} & \hat{a}_0(1 + \theta)|u_1|^2 + \hat{a}_N(1 + \theta)|u_N|^2 \\ & + \sum_{k=1}^{N-1} a_k |u_{k+1} - u_k|^2.\end{aligned} \tag{3.33}$$

Thus we see that for $\theta \neq -1$, the operator is positive-definite, and if $\theta = -1$, the nullspace consists only of the manifold spanned by the single function $u \equiv 1$.

When we go to m space variables, θ becomes a function defined on the boundary of the rectangular R. If θ is constant on each face of R (it can have different values on different faces), one shows easily that the operators $\bar{\nabla}_i\nabla_i$ commute among themselves. Thus if A is defined as in (3.6) with the operators defined as above, and $\theta \neq -1$ on any face, then we can also find an alternating-direction technique to solve $Au = y$ which converges at the same rate as before, i.e., in

$$O\left(\prod_{i=1}^{m} h_i \log(\min_i h_i)\right)$$

operations. If we are solving (3.1) with the boundary conditions (3.28) and the ratio $\beta(\bar{x})/\alpha(\bar{x})$ is bounded above and below, then we can proceed much as before. If β/α is bounded, $1 + \theta(\bar{x})$ is, for small h, approximately $h_i\alpha(\bar{x})/\beta(\bar{x})$ along the faces of R_h perpendicular to the i-axis. Looking for a moment at (3.33), we see that this has the same effect on $\varkappa_1$ and $\varkappa_2$ as does a change in the values of $\hat{a}_i$ on these boundaries. Thus if we let $1 + \theta_A = \max_i h_i \sup [\alpha(\bar{x})/\beta(\bar{x})]$ on each face of R, the change from Dirichlet to Robin boundary conditions is reflected only in the values of μ^0 and μ_0 in (3.5), and the convergence argument goes through as before with only minor changes. The values of the λ_i are no longer near $(\pi/l_i)^2$, of course, but are still near the corresponding (positive) minimum eigenvalues for the differential problem for small h.

The restrictions we must make on the Neumann problem are rather severe. If we could consider the iteration on the perpendicular complement of the nullspace $\mathfrak{N}$ of $\sum_i \bar{\nabla}_i\hat{a}_i\nabla_i$—just the set of identically constant net functions—all would be well, but we cannot, primarily because in general the solution is not to be found there. On $\mathfrak{N}$ the real part of the first-derivative operators can be large compared with the bounded zero-order term and this will in general cause divergence. In order that the error operator D_n in (3.9) have positive-definite real part for all h, it is necessary that the b_i be zero ($\hat{\nabla}_i$, skew for the Dirichlet case, has a nonzero real part here). We also require that either $\partial r/\partial u$ be zero for all u or somewhere positive for all u; in the first case, D_n is zero on $\mathfrak{N}$ and $\mathfrak{N}$ is a reducing subspace for all u; so we can speak of convergence on $\mathfrak{N}^\perp$, on which D_n is positive-definite. To get an estimate for the rate, it is necessary to impose a positive lower bound on $\partial f/\partial u$ if it is anywhere nonzero in order to insure a nonzero decay rate on $\mathfrak{N}$. Given these conditions, however, the analysis proceeds in the same fashion as before; for the interesting problem

$$\sum_{i=1}^{m} -\bar{\nabla}_i\hat{a}_i\nabla_i u = f, \tag{3.34}$$

where $\partial f/\partial u = 0$, one again establishes a rate $1/M_1$ times that for the alternating-direction iteration for $Au = y$, both considered on $\mathfrak{N}^\perp$.

For problems with Robin or Dirichlet data on some faces and Neumann on others, we must have $b_i = 0$ if both faces perpendicular to the x_i axes have Neumann data specified; otherwise, the analysis is similar to the Dirichlet case.

(iv) *A three-level normalized variant of the method.* A moment's reflection will show that the number of results which carry over in only slightly altered form from the "classical" iterative methods is rather large. Just as the application of the Čebyšev semi-iterative technique could be applied almost without change to the semi-explicit method, so also can most of the other results applicable to the older symmetric methods. In particular, the three-level "second-order Richardson" method [6], which yields the same result after any n iterations as the best n-parameter Čebyšev scheme and in which the round-off problem is eliminated, can be applied here. This process in a normalized version will probably prove to be the fastest of all the semi-explicit schemes; it is complicated by the fact that, like all three-level methods, the storage requirements for machine computation on large problems is quite severe.

Consider the Dirichlet problem for (3.21) of §3(ii), in the case when $a_i(x) \equiv a(x)$. In the differential case one easily verifies that

$$\sqrt{a}\,\frac{\partial^2}{\partial x^2}(u\sqrt{a}) = \frac{\partial}{\partial x}\left(a\,\frac{\partial u}{\partial x}\right) - \left(\sqrt{a}\,\frac{\partial^2}{\partial x^2}\sqrt{a}\right)u, \tag{3.35}$$

when a is bounded below and is sufficiently smooth. A similar result holds for the difference operators when a_i is defined appropriately.

Let

$$a_i(\bar{x}) = \{a(x_1, \cdots, x_i + h_i, \cdots, x_m)a(x_1, \cdots, x_i, \cdots, x_m)\}^{1/2}.$$

Then

$$\sqrt{a}\,\overline{\nabla}_i\nabla_i(\sqrt{a}\,u) = \overline{\nabla}_i a_i \nabla_i u + \sqrt{a}\,(\nabla_i\overline{\nabla}_i\sqrt{a})u. \tag{3.36}$$

The second term on the right is clearly of the form Bu, where B is a diagonal matrix whose norm can be bounded independent of the mesh if a is a smooth function. The computation of the a_i is more complicated than the usual arithmetic mean, but square roots of the values of a of each mesh point will be used in the operator on the left in (3.36) (which will become our iterator) and so must be calculated anyway. One shows easily that this definition of the a_i leads to an approximation of the differential equation locally second-order correct in the h_i.

We consider the iteration defined by

$$A(u^{n+1} - u^n) = \rho_1(L_h u^n - f) + \rho_2 A(u^n - u^{n-1}), \tag{3.37}$$

with L_k as in (3.21) and with

$$A = \sqrt{a}\left[-\sum_{i=1}^{m} \overline{\nabla}_i \nabla_i + \gamma\right]\sqrt{a}, \qquad \gamma = \frac{1}{2}\max_R\left(\frac{q}{a}\right).$$

The related eigenvalue problem is

$$L_h\psi_j = \lambda_j A\psi_j . \tag{3.38}$$

For convenience, let $A' = A - \gamma a$, $L_h' = L_h - q$, and $b = \sqrt{a}\sum_{i=1}^{m}(\overline{\nabla}_i\nabla_i\sqrt{a})$. Now γa, q, and b are all bounded diagonal matrices. Thus $C = a^{-1}q - a^{-1}b - \gamma$ is also a diagonal matrix whose norm is bounded by a quantity independent of the mesh size. Then (3.38) becomes

$$(A + q - b - \gamma a)\psi_j = \lambda_j A\psi_j , \tag{3.39}$$

or, for $\lambda_j \neq 1$,

$$A\psi_j = \frac{1}{\lambda_j - 1}\, aC\psi_j . \tag{3.40}$$

Then, letting $H = -\sum_{i=1}^{m}\overline{\nabla}_i\nabla_i + \gamma = a^{-1/2}Aa^{-1/2}$ and setting $\varphi_j = a^{1/2}\psi_j$, we obtain

$$H\varphi_j = \frac{1}{\lambda_j - 1}\, C\varphi_j = \mu_j\, C\varphi_j . \tag{3.41}$$

The problem in this form is quite similar to the eigenvalue problem for the vibrating membrane with variable density, except that C is *not* necessarily a positive-definite operator. We wish to estimate the μ_j in terms of the eigenvalues of H, which are known. The problem in this form is not too tractable, but we can transform again, setting $\mathfrak{X}_j = H^{1/2}\varphi_j$. Then

$$(\lambda_j - 1)\mathfrak{X}_j = H^{-1/2}CH^{-1/2}\mathfrak{X}_j \tag{3.42}$$

and

$$(\lambda_j - 1)^2\mathfrak{X}_j = H^{-1/2}CH^{-1}CH^{-1/2}\mathfrak{X}_j . \tag{3.43}$$

Now $H^{-1/2}CH^{-1}CH^{-1/2}$ is a positive-semidefinite operator, so we are in a position to use the Courant minimax theorem [7] to advantage. It is clear that

$$\begin{aligned}(u, H^{-1/2}CH^{-1}CH^{-1/2}u) &\leqq \|CH^{-1}C\|(u, H^{-1}u)\\ &\leqq \|C\|^2\|H^{-1}\|(u, H^{-1}u).\end{aligned} \tag{3.44}$$

Let α_k be the eigenvalues of H, arranged in increasing order; arrange the λ_k in decreasing order of $(\lambda_j - 1)^2$. Let $\mathfrak{M}$ denote any subspace of $\mathfrak{H}$, and let dim $\mathfrak{M}$ denote the dimension of $\mathfrak{M}$. Then by the minimax theorem,

$$(\lambda_j - 1)^2 = \inf \{\sup_{\substack{\|v\|=1 \\ u \in \mathfrak{M}}} (u, H^{-1/2}CH^{-1}CH^{-1/2}u) \mid \dim \mathfrak{M} = N - j - 1\}$$

$$\leqq \| C \|^2 \| H^{-1} \| \inf \{\sup_{\substack{\|u\|=1 \\ u \in \mathfrak{M}}} (u, H^{-1}u) \mid \dim \mathfrak{M} = N - j - 1\} \tag{3.45}$$

$$= \| C \|^2 \|H^{-1} \| \frac{1}{\alpha^j}.$$

So

$$| \lambda_j - 1 | \leqq \frac{\| C \| \| H^{-1} \|^{1/2}}{\sqrt{\alpha_j}} \leqq \frac{K}{\sqrt{\alpha_j}}, \tag{3.46}$$

where K is independent of the mesh size. Since the α_k tend, as $h \to 0$, to the corresponding eigenvalues of the differential operator $-\Delta + \gamma$, and these form an infinite sequence tending to infinity, it is clear that the number of λ_j which differ from 1 in absolute value by more than any fixed positive quantity remains bounded as $h \to 0$. These λ_j correspond to bounded μ_j in (3.41). Thus for small h, most of the λ_j are near 1. The few remaining are bounded above and below indpendent of h, since

$$\lambda_j = \frac{(\psi_j , L_h \psi_j)}{(\psi_j , A\psi_j)} \leqq \sup_{u \in \mathfrak{C}} \frac{(u, L_h u)}{(u, Au)} \leqq 1 + \sup_{u \in \mathfrak{C}} \frac{|(u, aCu)|}{(a, Au)}$$

$$\leqq 1 + \sup_{u \in \mathfrak{C}} \frac{|(u, Cu)|}{(u, Hu)} \leqq 1 + \frac{\| C \|}{\| H^{-1} \|}, \tag{3.47}$$

and

$$\frac{1}{\lambda_j} = \frac{(\psi_j , A\psi_j)}{(\psi_j , L_h \psi_j)} \leqq \sup_{u \in \mathfrak{C}} \frac{(u, Au)}{(u, L_h u)}$$

$$\leqq 1 + \sup_{u \in \mathfrak{C}} \frac{|(u, aCu)|}{(u, L_h u)} \leqq 1 + \frac{\| aC \|}{\| L_h^{-1} \|}; \tag{3.48}$$

we have seen that all the quantities appearing in these estimates have bounds independent of h. These bounds are not, in general, nearly so favorable as the ones found for the unnormalized procedure, but the estimates are quite crude. Much depends on the smoothness of the function $a(\bar{x})$.

We recast (3.37) in the form

$$A(u^{n+1} - \rho_2 u^n) = \rho_1(L_h u^n - f) + A(u^n - \rho_2 u^{n-1}). \tag{3.49}$$

If now we let $v^n = \sqrt{a}\, u^n$, $v = \sqrt{a}\, u$, we have

$$H(v^{n+1} - \rho_2 v^n) = \rho_1(a^{-1/2} L_h a^{-1/2} v^n - f) + H(v^n - v^{n-1}). \tag{3.50}$$

Now H can be inverted with alternating directions, and if we consider the

new scheme in which we approximate at each step the quantity $v^{n+1} - \rho_2 v^n$ by that obtained with one cycle of alternating-direction iteration, we obtain the new iterative procedure

$$B(u^{n+1} - \rho_2 u^n) = \rho_1(L_h u^n - f) + B(u^n - \rho_2 u^{n-1}), \tag{3.51}$$

where

$$B = \sqrt{a}(1 - \Lambda)^{-1} H \sqrt{a} = \sqrt{a}\, H^{1/2}(1 - \Lambda)^{-1} H^{1/2} \sqrt{a}$$

and Λ is the alternating-direction operator associated with H, as before. If we look at the eigenvalues β_j in

$$L_h \mathfrak{X}_j = \beta_j B \mathfrak{X}_j , \tag{3.52}$$

then the transformation $L_h^{1/2} \mathfrak{X}_j = \xi_j$ yields the equivalent form

$$\xi_j = \beta_j L_h^{-1/2} \sqrt{a}\, H^{1/2}(1 - \Lambda)^{-1} H^{1/2} \sqrt{a}\, L_n^{-1/2} \xi_j , \tag{3.53}$$

and a simple application of the minimax theorem yields the estimate

$$(1 - \zeta)\lambda_j \leqq \beta_j \leqq (1 + \zeta)\lambda_j , \tag{3.54}$$

where $\zeta = \| \Lambda \|$, as before. Thus as $h \to 0$, the number of β_j lying outside $(1 - \zeta - \epsilon,\ 1 + \zeta + \epsilon)$ remains bounded for any fixed $\epsilon > 0$, and $\beta_0 \leqq \beta_i \leqq \beta^0$, where β_0 and β^0 do not depend on h.

The error $e^n = B^{1/2}(u^n - u)$ satisfies

$$(e^{n+1} - \rho_2 e^n) = \rho_1 B^{-1/2} L_h B^{-1/2} e^n + (e^n - \rho_2 e^{n-1}). \tag{3.55}$$

The eigenvalues of $B^{-1/2} L_h B^{-1/2}$ are just β_j, and the eigenfunctions are $\overline{\mathfrak{X}}_j = B^{1/2} \mathfrak{X}_j$. Since $B^{-1/2} L_h B^{-1/2}$ is self-adjoint, the $\overline{\mathfrak{X}}_j$ are complete and can be chosen as an orthonormal set. Thus we can write

$$e^n = \sum d_i^n \overline{\mathfrak{X}}_i , \tag{3.56}$$

and $\| e^n \| = \sum | d_i^n |^2 = \| u^n - u \|_B$. Then, using (3.55), we obtain

$$d_i^{n+1} - \rho_2 d_i^n = \rho_1 \beta_i d_i^n + (d_i^n - \rho_2 d_i^{n-1}). \tag{3.57}$$

This is an ordinary difference equation with constant coefficients for each i; the characteristic polynomial is

$$x^2 - (1 + \rho_2 + \beta_i \rho_1)x + \rho_2 = 0. \tag{3.58}$$

Let $\gamma_1(\beta_i)$ and $\gamma_2(\beta_i)$ be the two roots of (3.43). We wish to minimize

$$r = \max_{\beta_0 \leqq \beta \leqq \beta^0} \{ \max_{j=1,2} | \gamma_i(\beta) | \}$$

for ρ_1 and ρ_2 real. This problem was investigated and solved by Frankel [6] in connection with the second-order Richardson technique. The best choice of ρ_1 and ρ_2 is that couple which makes the roots complex for

$\beta_0 \leqq \beta \leqq \beta^0$. If $\varkappa = \beta_0/\beta^0$, ρ_1 and ρ_2 are given by

$$\rho_2 = 1 + \frac{8\varkappa}{(1-\varkappa)^2} - \sqrt{\left(1 + \frac{8\varkappa}{(1-\varkappa)^2}\right)^2 - 1},$$
$$\rho_1 = -\frac{1}{\beta^0}\left(\frac{2(1+\rho_2)}{1+\varkappa}\right), \tag{3.59}$$

as may easily be verified. Then $r = \sqrt{\rho_2} \sim 1 - 2\sqrt{\varkappa}$, as $\varkappa \to 0$. The convergence rate in terms of the ratio of least to greatest eigenvalue is thus like the Čebyšev process discussed earlier. It is in fact easy to see that one could do the three-level iteration without normalization and obtain the same convergence estimate as in §3(ii); alternately, normalization could be used in the two-level process, either using a fixed ρ or with a Čebyšev sequence. The latter especially might be used to advantage, since a few ρ's in the vicinity of 1 would effectively remove all the high-frequency components of the error.

We have remaining the practical problem of determining ρ_1 and ρ_2, which in turn, by (3.59) is equivalent to determination of β_0 and β^0. Probably the simplest method is the classical one; one runs two iterations for a few steps each with $\rho_2 = 0$, one with $-\rho_1$ large enough to cause divergence, the other with $-\rho_1$ small. In the first case, the limiting value of $\dfrac{\| u^{n+1} \|}{\| u^n \|}$ is $-\rho_1\beta_0 - 1$; in the other, $1 + \rho_1\beta_0$. It is clearly preferable to overestimate β^0 and underestimate β_0 than vice versa.

4. Results of numerical experiments. The semi-explicit iterative technique was tested numerically for the Dirichlet problem for

$$\nabla \cdot a \nabla u = f \tag{4.1}$$

on the unit cube in three dimensions, using $h = \frac{1}{16}$. The Douglas-Brian alternating-direction technique was used for the inversion of the iterator, as descussed in §3(i). Comparison was made with the ordinary Douglas-Brian [2], [4] alternating-direction method and with point successive overrelaxation [11] for Poisson's equation and for (4.1). Several a's were tried, some smooth and some generated on the net with a random-number generator employing a rectangular distribution in $[\mu_0, 2 - \mu_0]$. The ratio μ^0/μ_0 was of the order of 100 for most of the problems run. The unnormalized, constant ρ iteration proceeded more slowly, of course, than the parent iterative process, but much more rapidly than the μ^0/μ_0 factor predicted in §1. The Douglas-Brian procedure reduced the norm of the error by a factor of 2×10^7 in 21 iterations (one iteration was counted as one triple sweep of alternating-direction iteration) when iterating Poisson's equation. The average convergence rate for the semiexplicit method for

(4.1) after an equal amount of computation was about one-fourth this fast, the error norm descreasing by a factor of 100 in 21 iterations. When the Čebyšev process with a sequence of ten ρ_n was used, this figure increased to about 2×10^3.

The square-root normalizing scheme was not tested, but a nonsymmetric variant was, in the form

$$A(u^{n+1} - u^n) = \rho a^{-1}(L_h u^n - f). \tag{4.2}$$

Convergence rates were drastically increased, though the results were somewhat erratic and seemed to depend rather strongly on the form of the function a—a not totally unexpected result when one examines the skew part of the operator on the left in (4.2). It is the author's opinion that the normalization scheme proposed in §3 (iv) would yield much superior results. A further increase in speed was noted in some cases when the iterate u^n was advanced after each triple sweep instead of each cycle. The error reductions in 21 triple sweeps here varied from 10^4 to better than 10^8.

When the Douglas-Brian technique was applied directly to the problem (4.1), rates were observed which were quite comparable to the normalized semi-explicit method; sometimes slightly faster, often slightly slower, but never very different. Problems which appeared difficult for one also were slow with the other.

The test problem was sufficiently small that successive overrelaxation [11] took less computer time than either alternating directions or the semi-explicit technique, though about two and one-half times as many iteration steps were required (again counting an iteration step for the other methods as one triple sweep of alternating direction) to reduce the norm an amount equivalent to the other methods. The fact that three-dimensional problems will tend to be small for some time to come because of machine limitations makes the simpler overrelaxation technique relatively more attractive than the asymptotically faster method outlined in this paper. With the faster machines of the future, however, on large three-dimensional problems and for present two-dimensional problems the technique should yield very satisfactory results.

5. Acknowledgment. The author wishes to express thanks to D. N. Turner, who prepared the experimental programs and conducted most of the experiments.

REFERENCES

[1] L. Bers, *On mildly nonlinear partial difference equations of elliptic type*, J. Res. Nat. Bur. Standards, 51 (1953), p. 229–236.

[2] P. L. T. Brian, *A finite difference method of high-order accuracy for the solution of three-dimensional heat conduction problems*, A. I. Ch. E. J., (1961), pp. 367–370.

[3] E. G. D'JAKANOV, *On an iterative method for the solution of a system of finite-difference equations*, Dokl. Akad. Nauk SSSR, 138 (1961), pp. 522–525.

[4] J. DOUGLAS, *Alternating direction methods for three space variables*, Numer. Math., 4 (1962), p. 41–63.

[5] J. DOUGLAS AND J. E. GUNN, *A general analysis of alternating direction methods, part II: elliptic problems*, to appear.

[6] S. P. FRANKEL, *Convergence rates of iterative treatments of partial differential equations*, MTAC, 4 (1950), pp. 65–75.

[7] P. R. HALMOS, *Finite Dimensional Vector Spaces*, Van Nostrand, Princeton, 1958, p. 81.

[8] J. E. GUNN, *The numerical solution of* $\nabla \cdot a \nabla u = f$ *by a semi-explicit alternating direction iterative method*, Numer. Math., 6 (1964), pp. 181–184.

[9] W. MARKOV, *Über Polynome, die in einem gegeben Intervalle moglichst wenig von Null abweichen*, Math. Ann., 77 (1916), pp. 213–258.

[10] D. W. PEACEMAN AND H. H. RACHFORD, JR., *The numerical solution of parabolic and elliptic differential equations*, J. Soc. Indust. Appl. Math., 3 (1953), pp. 28–41.

[11] D. YOUNG, *Iterative methods for solving partial difference equations of elliptic type*, Trans. Amer. Math. Soc., 76 (1954), pp. 92–111.

[12] ———, *On Richardson's method for solving linear systems with positive definite matrices*, J. Math. and Phys., 32 (1954), pp. 243–255.